Detection of incorrect and inappropriateImagefrom Tweets in Social Network

Prof. Priyanka A. Jalan
Dept. of Computer Engineering
Bapurao Deshmukh College of Engg. Sevagram
Wardha, India
priyanka9.jalan@gmail.com

Rahul R. Sawarkar
Dept. of Computer Engineering
Bapurao Deshmukh College of Engineering, Sewagram
Wardha, India
rahul180216@gmail.com

Shubham M. Saral
Dept. of Computer Engineering
Bapurao Deshmukh College of Engineering, Sewagram
Wardha, India
shubham.saral04@gmail.com

Ashwini R. Khobragade
Dept. of Computer Engineering
Bapurao Deshmukh College of Engineering, Sewagram
Wardha, India
ashwinikhobragade31@gmail.com

***Abstract*—**Digital imaging has grown to become the prevalent technology for creating, processing, and storing digital memory and proof. Though this technology brings many leverage, it can be used as a ambiguous tool for covering details and evidences. This is because today digital images can be tampered in such supremacy that forgery cannot be find visually. In fact, the immunity concern of digital content has arisen a long time ago and different methods to verify the efficiency of digital images have been developed. Digital images offer many features for forgery detection algorithm to take precedence of specifically the color and brightness of individual pixels as well as an image's resolution and format. These properties grant for analysis and similarity between the significance of digital forgeries in an attempt to develop an algorithm for detecting image tampering. This paper presents a technique for image copy or move image forgery detection using Radix Sort, FasterK-means clustering algorithm & DCT

Keywords*—Faster K-means clustering algorithm, DCT,Image forgery , Image forgery detection , Radix Sort .*

*****

I. Introduction

Now a days social media has been the important part of one's life from education to electronic mail and business tool. Social media plays a vital role in transforming people's lifestyle. It includes social networking sites such as Twitter, Facebook and blogs where people can easily connect with each other. As per the survey of previous research, 90% of college students use social networks. Technology has shown the rapid development by introducing small communication devices and we can use them anytime anywhere. So social crimes are also increasing. However we present a method to detect the spam users, text and images by using different algorithms like Faster K-means clustering Algorithm and DCT

A Challenge

As far as the security is concerned in social media some of the challenges also comes up and motivate us to resolve them and work upon them . Some of the challenges has been discuss below.

1. Sentence detection:

While processing the natural language, deciding the beginning and the end of the sentences is one of the problems to be address. This process is known as Sentence Boundary Disambiguation(SBD) or simply sentence breaking. The techniques we used to detect the sentences in the given text depends on the language of the text.

2. **Mnemonics:**

Mnemonics aims to translate information into a form that the brain can retain better than its original form. Even the process of merely learning this conversion might already aid in the transfer of the information to long term memory.

3**. Similarities**:

Similarities are the metric defined over a set of documents or terms where the idea of distance between them is based on the likeness of their meaning or semantic content. This are mathematical tools used to estimate the strength of the semantic relationship.

II. Related Work

In most of the other approaches the suspicious image is divided into overlapping blocks. The aim is to detect blocks that are copied and moved. The copied region will contain overlapping blocks. The distance between each duplicate block pair will be similar since each block is carried with equal amount of shift. The next step would be extracting features form these blocks, which will give similar values for matching block. Different features can be used to perform theimage block. These blocks are vectorized and arranged in a matrix and the vectors are Radix sorted for later detection. The computational time calculate by number of blocks, sorting techniques and the number of feature. In this an image size of size P x Q, it is divided into (P-b+1) (Q-b+1) overlapping blocks of size b × b. The blocks are than sorted. Vectors related to blocks of matching content would be similar to each other in the list, so that same regions could be easily detected.

A. C. Popescu et. al.,[4] state that PCA is efficient to extract the image features. The method to produce each feature vector is called principle component analysis. The values are obtained by using the theorems of covariance matrix, eigenvectors and linear basis for each image block with the initial conditions of zero-mean. Then the vectors coefficients of each block are stored in a matrix S. These coefficients are then sorted lexicographically and the duplicated regions will be revealed by considering the offset of all pairs whose distances in S less than a specify threshold.

Ashima Gupta et. al.,[2] worked on an approach that can detect forged JPEG images and further locate the tampered parts, by examining the double quantization effect hidden among the Discrete Cosine Transform (DCT) coefficients. The image is divided into overlapping blocks (16x16) for feature extraction. Authors have used DCT coefficients for feature extraction and then find the matching blocks in the image.

Zhang et. al., [5] proposed an approach for detecting copymove forgery detection in digital images. Authors used Discrete Wavelet Transform (DWT) and divided lowfrequency band into four non-overlapping sub-images and phase correlation is used to compute the spatial offset between the copy-move regions. Then, they applied pixel matching for detecting the duplicate region. This algorithm works well in the highly compressed image. This is an effective algorithm with lower computational time compared with other algorithms.

Xiao Bing Kang et. al., [6] introduce an algorithm named Singular Value Decomposition (SVD) was used to extract the algebraic and geometric features from small overlapping image blocks and to produce a singular value feature vectors which are saved in a matrix. This matrix is then reduced rank by reduced-rank approximation before detecting the similarity of vectors.

M. K. Bashar et. al.,[7] given Kernel Principle Component Analysis (KPCA) or wavelet transform to extract the features of the small blocks split from a given image which are then lexicographically sort to suggest the similarity of corresponding blocks. The paper proposes algorithms to detect forged areas with translation, flip and rotation based on the global. The results also examine cases of addition noise and lossy JPEG compression. KPCA is the best in case of noisy and rotation of any degree compared with PCA and wavelet based.

Kakar and Sudha et. al.,[8] developed a new technique based on transform-invariant features which detecting copy-paste forgeries but need some post processing based on the MPEG image signature tools. Feature matching that uses the inherent constraints in matched feature pairs so as to improve the detection of tampered regions is used which results in a feature matching

Sutthiwan et. al.,[10] presented a method for passive-blind colour image forgery detection which is a combination of image features extracted from image luminance by applying a rake – transform and from image by using edge statistics.

Huang et al.,[11] proposed a copy move forgery detection method based on Scale Invariant Feature Transform (SIFT) descriptors. After extracting the descriptors of different regions, they match them with each other to find manipulated area in images.

Fridrich et. al.,[12] used Discrete Cosine Transform (2DDCT). They use lexicographic sorting after extracting 2DDCT coefficients of each block in an image. Then find the equivalent distance between each block. If distance is less than the image is forged

Ghorbani et. al.,[13] proposed DWT-DCT (QCD)-based copymove image forgery detection in 2011. Authors used DWT and resolved the image into sub-bands and then performed DCT-QCD (quantization coefficient decomposition) in row vectors to reduce vector length. After lexicographically sorting the row vectors, shift vector is computed. Finally, the shift vector is compared with threshold and the forged region is highlighted.

Lin et. al.,[14] proposed an combined technique for splicing and copy-move image forgery detection in 2011. First, the authors converted an image into the grey. For splicing detection, the image is divided into sub-blocks and DCT is used for feature extraction and SURF is used for copy-move detection. The algorithm works in both splicing and copy move image forgery detection.

Leida Li et. al.,[15] this paper presents a method for detecting image forgery based on circular pattern matching. The tampered image is filtered and divided into circular blocks. Using Polar Harmonic Transform (PHT) rotation and scaling feature is extracted from each block. The feature vectors are lexicographically sorted and the manipulated regions are detected by finding the similar block pairs after applying postprocessing.

III. PROPOSED WORK

An image extraction use for blurring the image so the user can only see the blurred image and will not be visible to anyone else. For the image extraction Faster K-means algorithm using Radix sort and DCT i.e., Discrete Cosine Transformation has been used. The primary purpose of copy-move forgery is to detect similar regions in an image the duplicated regions are unknown both in size and shape. It is not easy to compare every pairs pixel by pixel, as well as it guides to higher computational complexity. In order to make efficient forgery detection an image window is used. Some appropriate and robust features are extracted from the image window, an efficient features extraction not only represent the whole image window, but also lower the dimension of feature vector, and due to less dimension of feature matrix, forgery detection algorithm has less computational complexity.

3.1 Pre-processing of the image

The proposed method operates in the luminance domain.Therefore, the colour image is first converted into gray scale by

$I = 0.228R + 0.587G + 0.114B$ (1)

where R, G, B denote the red, green and blue components of the image.

There are a number of types of image file These include the following.

PNG:

PNG is a lossless compression type. It is often used where the graphic might be changed by another person or there the image contains layers of graphics that need to be kept separate from each other. It is high quality

JPEG:

JPEG is often used for digital camera images because it has a fairly small file size for the quality that it displays. JPEG is a lossy format that offers a higher compression rate than PNG in the trade-off for quality.

GIF:

GIF compresses images to a maximum 8-bit color depth, making it unsuitable for high quality photographs. GIF is often used where transparency is needed on the graphic. GIF can also be store simple animated images.

3.2Faster K-Means Clustering

- K-means depends mainly on distance calculation between all data points ans the centers, the time cost will be high when the size of the dataset is large(for example more than 500million of points). This two stage algorithm to reduce the time cost of distance calculation for huge datasets.
- The sirst stage is a fast distance calculation for huge calculation using only a small portion of data to produce the best possible locations of the centers.
- The second stage is a slow distance calculation in which the initial centers used are taken from the first stage.
- The time cost of the distance calculation for the training data chosen. The time cost of the distance calculation for the fast stage is very low due to the small size of the training data chosen.
- The time cost of the distance calculation for the slow stage is also minimized due to small number of iterations. Different initial locations of the cluster have been used during the test of the proposed algorithm.
- The method of **Vector Quantization,** originally from signal processing, that is popular for cluster analysis in data mining.Faster K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean

Description: Given a set of observation (x1,x2……xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observation into k(<n) sets **S**={s1,s2…..sk} so as to minimize the within cluster sum of squares

$$\text{args min}\sum_{i=1}^{k} 1/2|s\,\textstyle\sum_{x,y€s}^{n} ||\mathrm{x}-\mathrm{y}||^{2}$$

Vector quantization:

(VQ) is a classical quantization technique from signalprocessing that allows the modeling of probability density functions by the distribution of prototype vectors. It was originally used for datacompression. It works by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them. Each group is represented by its centroid point, as in k-means and some other clustering algorithms.
A simple training algorithm for vector quantization is [1]:

1. Pick a sample point at random
2. Move the nearest quantization vector centroid towards this sample point, by a small fraction of the distance
3. Repeat

A more sophisticated algorithm reduces the bias in the density matching estimation, and ensures that all points are used, by including an extra sensitivity parameter :

1. Increase each centroid's sensitivity by a small amount
2. Pick a sample point at random
3. For each quantization vector centroid , let denote the distance of and
4. Find the centroid for which is the smallest
5. Move towards by a small fraction of the distance
6. Set to zero
7. Repeat

Applications:

1. Vector quantization is used for lossy data compression, lossy data correction, pattern recognition, density estimation and clustering.
2. Lossy data correction, or prediction, is used to recover data missing from some dimensions. It is done by finding the nearest group with the data dimensions available, then predicting the result based on the values for the missing dimensions, assuming that they will have the same value as the group's centroid.

3.3 DCT (Discrete Cosine Transformation)

- Discrete cosine transform. A descrete cosine transform (DCT) expresses a finite sequence of data points in terms of a sum of cosine functions oscillating at different frequencies. The most common variant of discrete cosine transform is the type-II DCT, which is calledsimply “the DCT”.
- The DCT, and in particular the DCT-II, is often used in signal and image processing, especially for lossy compression, because it has a strong “energy compaction” property.

- The DCT is at the heart of the international standard lossy image compression algorithm known as JPEG.
- The DCT is used in jpeg image compression, mjpeg , mpeg, DV , Daala and theora video compression.
- A new variety of fast algorithms are also developed to reduce the computational complexity of implementing DCT.

What is image transform?

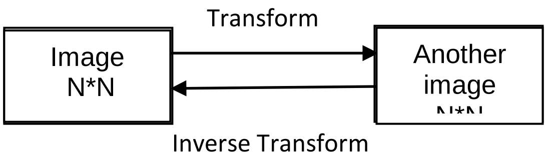


Figure 3.1: Working of DCT

DCTFormula:

$$\mathrm{DCT}(i,j) = \frac{1}{\sqrt{2N}} C(i)\,C(j) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} \mathrm{pixel}(x,y) \,\mathrm{COS}\left[\frac{(2x+1)i\pi}{2N}\right] \mathrm{COS}\left[\frac{(2y+1)j\pi}{2N}\right]$$

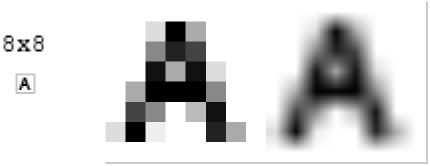
$$C(x) = \frac{1}{\sqrt{2}} \text{ if x is 0, else 1 if } x > 0$$

Application:

- The DCT, and in particular the DCT-II , is often used in signal and image processing, especially for lossy compression, because it has a strong “energy compaction” property. In typical applications, most of the signal information tends to be concentrated in a few low-frequency componets of the DCT.
- For strongly correlated Markov processes , the DCT can approach the compaction efficiency.

Example of DCT

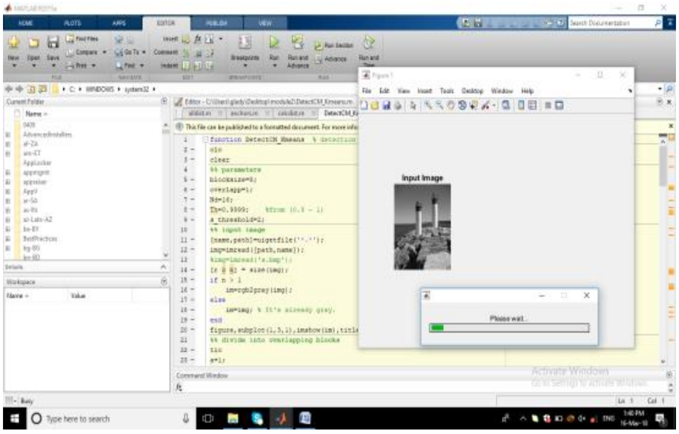
Consider this 8x8 grayscale image of capital letter A.



Original size, scaled 10x (nearest neighbor), scaled 10x (bilinear).

IV. EXPERIMENTAL REVIEW

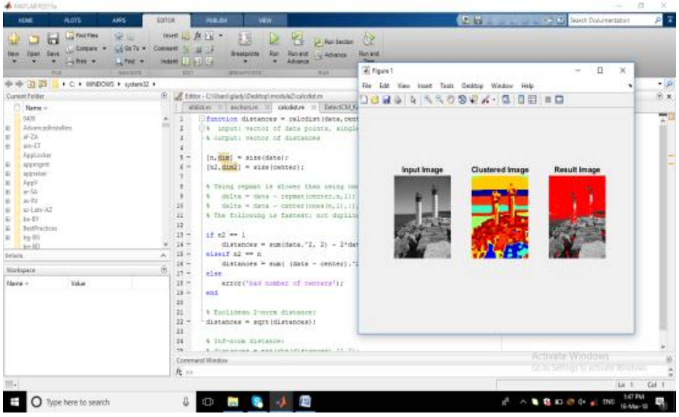
4.1. Snapshots



```
function DetectCM_Kmeans  % detection
clc
clear
%% parameters
blocksize=8;
overlapp=1;
Nd=16;
Th=0.9999;    %from (0.9 - 1)
s_threshold=2;
%% input image
[name,path]=uigetfile('*.*');
img=imread([path,name]);
%img=imread('s.bmp');
[r c k] = size(img);
if k > 1
    im=rgb2gray(img);
else
    im=img; % It's already gray.
end
figure,subplot(1,3,1),imshow(im),title
%% divide into overlapping blocks
tic
a=1;
```

Snapshot 4.1.1

The aim of pre-processing is an improvement of the image data that suppress unwanted distortions or enhances some image features important for further processing.



```
function distances = calcdist(data,cent
%  input: vector of data points, single
% output: vector of distances

[n,dim] = size(data);
[n2,dim2] = size(center);

% Using repmat is slower than using one
%   delta = data - repmat(center,n,1);
%   delta = data - center(ones(n,1),:);
% The following is fastest: not duplic

if n2 == 1
    distances = sum(data.^2, 2) - 2*dat
elseif n2 == n
    distances = sum( (data - center).^2
else
    error('bad number of centers');
end

% Euclidean 2-norm distance:
distances = sqrt(distances);

% Inf-norm distance:
```

Snapshot 4.1.2

After getting processed image gray scale images use just one channel of color, that normally is necessary just 8 bit to be represented

V. CONCLUSION

The copy-move forgery detection is one of the emerging problems in the field of digital image forensics. In the last decade many forgery detection techniques have been proposed. An attempt is made to bring in various potential algorithms that signify improvement in image authentication techniques. The techniques which have been developed till now are mostly cable of detecting the forgery and only a few can localize the tampered area. There are many drawbacks with the presently available technologies. Firstly all systems require human interpretation and thus cannot be automated.

REFERENCES

[1] Mohd Dilshad Ansari, S. P. Ghrera & Vipin Tyagi 2014 Pixel-Based Image Forgery Detection: A Review. IETE Journal of Education.

[2] Ashima Gupta, Nisheeth Saxena, S.K Vasistha 2013 Detecting Copy Move using DCT, International Journal of Scientific and Research Publications.

[3] Vivek Kumar Singh and R.C. Tripathi 2011 Fast and Efficient Region Duplication Detection in Digital Images

Using Sub-Blocking Method. International Journal of Advanced Science and Technology

[4] A. C. Popescu, and H. Farid 2004 Exposing digital forgeries by detecting duplicated image regions. Dept. Comput. Sci., Dartmouth College.

[5] J. Zhang, Z. Feng, and Y. Su 2008 A new approach for detecting copy-move forgery in digital images. In IEEE International Conference on Communication Systems, China.

[6] XiaoBing KANG, ShengMin WEI 2008 Identifying Tampered Regions Using Singular Value Decomposition in Digital Image Forensics. IEEE International Conference on Computer Science and Software Engineering, Wuhan, Hubei.

[7] M. K. Bashar, K. Noda, N. Ohnishi, and K. Mori 2010 Exploring Duplicated Regions in Natural Images", IEEE Transactions on Image Processing.

[8] P. Kakar and N. Sudha 2012Exposing post processed copy-paste forgeries through transform-invariant features. IEEE Trans Inf Forensics Security.

[9] Muhammad, M. Hussain and G. Bebis 2012 Passive copy move image forgery detection using undecimated dyadic wavelet transform. Digital Investigation.

[10] P. Sutthiwan, Y. Q. Shi, S. Wei and N. Tian-Tsong 2010 Rake transform and edge statistics for image forgery detection. Proc. IEEE International conference on multimedia and Expo.

[11] Huang H, Guo W, Zhang Y 2008 Detection of copy-move forgery in digital images using SIFT algorithm. In: Proc. IEEE Pacific-Asia workshop on Computational Intelligence and Industrial Application.

[12] Fridrich J, Soukal D, Lukas J 2003 Detection of copy-move forgery in digital images. In: Proceedings of Digital Forensic Research Workshop.